DOCUMENT RESUME

ED 194 630 TM 800 762

AUTHOR Ladas, Harold
TITLE The Power of An Experiment: Potential Vs. Actual.
PUB DATE Apr 80
NOTE 31p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Control Groups; *Educational Research; Higher Education; *Quasiexperimental Design; *Research Design; *Research Methodology; *Research Problems
IDENTIFIERS Power (Statistics)

ABSTRACT

A relationship between experimental and statistical controls is considered in an effort to show how this relationship affects the power of an experiment. Statistical rather than experimental controls are often used to reduce the effect of subjects at the extremes of individual difference. This results in attrition of actual experimental subjects. Questionnaire data from two experiments was analyzed for subject attrition, although they were not specifically designed for this purpose. In an experiment on cues and notetaking using 54 graduate students as subjects, there was a failure to find two strong expected treatment effects (cues over no cues and notes over no notes). In the second experiment on lecture vs. reading with and without notes using 58 college juniors and seniors, the treatment group which heard the lecture and took notes had 10 non-subjects and only six subjects. In both experiments, substantial power loss from subject attrition was found. Recommendations are made for conducting experiments and summarizing research. Inadequate actual power of experiments reporting no difference may be a major reason for pessimistic research summaries. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED194630

The Power of An Experiment: Potential Vs. Actual

Harold Ladas

Department of Educational Foundations

Hunter College of the City University of New York

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. Ladas

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the American Educational Research Association Meeting, Boston, Mass. , April 9, 1980.

TM800762

Running Head: The Power of An Experiment

Abstract

A relationship between experimental and statistical controls is considered in an effort to show how this relationship affects the power of an experiment. Statistical rather than experimental controls are often used to reduce the effect of subjects at the extremes of individual difference. This results in attrition of actual experimental subjects. Questionnaire data from two experiments was analyzed for subject attrition. Substantial power loss from subject attrition was found. Recommendations are made for conducting experiments and summarizing research. Inadequate actual power of experiments reporting no difference may be a major reason for pessimistic research summaries.

Charters and Jones (1973) describe the risk of appraising non-events in program evaluation. This paper illustrates what may be a comparable problem in educational and psychological research: conducting and reporting non-experiments. The source of the error is identified as a loss of power which occurs at the interface between experimental design and statistical inference; recommendations for correction are made.

The Power of An Experiment

The power of a statistical test is defined as $1 - \beta$ (beta), i.e., the probability of rejecting the null hypothesis (when it is false and should indeed be rejected). Assuming that the statistical test corresponds to the realities of the experiment, power represents the probability of finding a difference between treatment groups when it does exist. Power is determined by α (alpha) level, the sample size n, and the magnitude of the effect in the population. As each of these parameters increases, power increases. Cohen (1962) suggested these arbitrary but reasonable descriptions of effect size. A small difference between population means is defined as $.25\sigma$ (sigma), a medium difference is defined $.5\sigma$ (sigma), and a large difference is 1σ (sigma) where σ (sigma) is the standard deviation of the population.

Experience has shown that most of the effect sizes

found in teaching and learning experiments are not large. Gage (1978) believes this occurs because many variables are involved and no single variable accounts for very much of the variance. The research on leadership which in many ways is analogous to teaching supports this view. Although many variables are correlated with effective leadership (eg., Stogdill, 1974, pp. 74-81), no single trait or behavior alone accounts for a large portion of the variance.

There is another reason why we seldom obtain large effects. Many such experiments would be regarded as unacceptable for journal publications. Consider these examples. Treatment X_1 consists of presenting material at an inaudible level. Treatment X_2 consists of presenting the identical material on an audible level. Both treatments are followed by the same four option, multiple choice, edumetric (criterion referenced) test. Treatment group X_1 scores at the chance level, 25%. Treatment group X_2 scores 75%. The difference between a chance score of 25% and 75% would result in a very large effect size. In Treatment Y_1 all the students are asleep (a very low arousal state), whereas, in Y_2 they are optimumly aroused. Granted that sleeping subjects can learn something, say 10%, there should be extremely large ef-

fect sizes. Since the charge is often made that we know little or nothing about teaching and learning, it should be noted that we "know" somethings so well that an experiment to prove the point would be regarded as trivial. Ladas (1977) performs a number of these experiments in vacuo to illustrate some implications for research and teaching.

Power calculations are usually based on the number of subjects collected or selected for an experiment and who turn in a criterion response. Here is an example of a power calculation based on α (alpha) = .05 an effect size of σ (sigma) = .50 and an N = 44. Assuming that a directional hypothesis was stated and given that two groups are involved:

$$\delta = \frac{|u_1 - u_2|}{\sigma\sqrt{\frac{1}{n_1}\ \frac{1}{n_2}}} = \frac{.5}{\sqrt{.09}} = 1.667$$

f = 22 + 22 -2 = 42

Power = .46 (from Table eg. Guenther 1965 p. 315)

If no hypothesis was stated, a two tailed test would be used and the power is .34.

Brewer (1972) calculated the power of 373 statistical test results which were reported as significant in AERJ. The combined average power of all 373 tests was

found to be .17, .58 and .78 assuming small, medium, and large effect sizes respectively (p. 393). Similar power values were obtained by Brewer (1972) analyzing articles in the Journal of Research in Science Teaching and the Research Quarterly, and by Cohen (1962) in his survey of the Journal of Abnormal and Social Psychology. Interpreting the results of these analyses, Brewer points out that over two-thirds of the tests had a power less than .50 for small and medium effects (p. 393). And Brewer believes that "researchers are testing for effect sizes which are smaller than the small size described by Cohen (1969)" (p. 394). Researchers would have a better chance of rejecting false null hypotheses by flipping a coin!

Particularly when the result of an experiment is "no difference" it is relevant to ask whether the experiment had a reasonable probability of detecting a difference if there was one. This failure to report power is analagous to a biologist failing to find bacteriophages and thus concluding that they do not exist, without stating the magnification or resolving power of the microscope he used.

The Actual Power of An Experiment

The introduction discussed what could be defined as the potential power of an experiment. The concept of potential power assumes that all subjects selected or collected, who produce a criterion statistic are active, viable participants. There are strong theoretical and practical reasons for doubting this assumption.

The theoretical reasons revolve around individual differences. Since the time of Galton it has been known that individual differences are normally distributed. What is relevant to this article, is the range of those differences. Carroll (1963) pointed out that aptitude (I.Q.) can be equated to learning rate and that learning rates differ by a factor of at least four.

For example, let us suppose that the mean learning rate for completing an experimental treatment is 25 minutes. If Carroll's work applies, some subjects will complete the treatment in 10 minutes while others take 40 minutes. The problems lie in the extremes of the range. For subjects who will complete in 10 minutes, there is the problem of what to do with them so they don't disrupt other subjects who are still working. Regarding subjects who will complete in 40 minutes, there is the problem of controlling for experimental subjects who have

not finished. Often the experimenter arbitrarily ends the experiment before all subjects have actually completed; the result is subject attrition.

Subject attrition is defined as subjects selected or collected for an experiment who do not meet experimental control specifications, eg. are unable or unwilling. Achievement levels also differ in a population of subjects. Trow (1956) states that the achievement span in elementary school is six years. That means that in the third grade some subjects are doing first grade work, others are doing sixth grade work. Achievement spans of 14-16 years are found in secondary schools. (Jencks, Smith, Acland, Bane, Cohen, Gintis, Heyns, Michelson, 1972).

At the college level, the achievement span situation is unclear. Admissions requirements and self-selection are assumed to truncate the lower part of the population distribution, but open admissions reaches deeply into this population. For example, 70% of the SEEK students at Hunter College read below ninth grade level.

As Confucious observed, "Men are born pretty much alike, but through their habits, they gradually grow further and further apart from each other." From a research point of view, this often means that on the one hand, some

subjects already know part or all of our treatment material, have superordinate organizers to immediately understand it or can answer some or many of the criterion items from cues, while on the other hand, some subjects do not have the prerequisite behavior or skills necessary to begin the treatment material. The research consequence of these differences is an inevitable and unpredictable degree of subject attrition.

The position on the normal distribution curve (i.e., percentile rank) of an individual on one trait is not highly predictive of his position regarding other traits. For example, an individual might be high in motivation, average in aptitude and low in ability to follow directions. Some individuals are high in math skills and low in language ability and vice versa. From the viewpoint of research, this means that there can be subject attrition for each aptitude or trait relevant to an experimental control, and that this attrition is unpredictable and cumulative for any given subject sample.

Aptitude treatment interactions (Note 1) or more broadly, attribute treatment interactions (Note 2) partly resolve this problem, but (ATIs) only do so for a few experimental variables at a time. There can still be substantial subject attrition at each of the controlled variables.

Of the subjects who volunteer or are collected for an experiment, some cannot or will not conform to the requirements of the experimental controls. These are non-subjects. Experimenters are often unaware of non-subject's non-compliance or non-conformance to specifications, so these non-subjects masquerade as bona fide subjects. The actual number of experimental subjects is the potential number collected minus the number of non-subjects. The actual power of the experiment is the power estimate based on the actual N of subjects. We shall see in the experiments described below that the actual power is substantially lower than the potential power.

Some Experimental Evidence

The following two experiments were not specifically designed to determine subject attrition. The failure of the investigators to find two strong expected treatment effects (cues over no cues and notes over no notes) led to a reanalysis of the actual power of the first experiment. A previous experiment with a significant difference favoring lecture over reading and note taking over not taking notes was then reanalyzed and also found suspect.

An Experiment on Cues and Notes

Fifty-four subjects from four graduate classes in

special education were randomly assigned to four orthagonal treatment groups: cues—no cues, notes—no notes. The treatment consisted of a 3,600 word lecture on psycholinguistics which was read at the rate of 100 wpm. Subjects in the no cue groups sat on one side of a barrier so they could not see the cue to take notes. (A 3" x 5" card with a green circle was exposed by an assistant to indicate when to take notes). The cue to take notes was given when the lecturer presented material directly to the 50 question criterion test; about 30% of the lecture material was cued for note taking. Since subjects take notes at about 20 wpm it was assumed that subjects in the "cues given" condition would have a distinct advantage.

The criterion test had a split half reliability of .76. It was given after review of notes which were returned to subjects for 10 minutes one week after the treatment. The notes were collected before the criterion test was administered.

The investigators expected strong positive effects both from cues and from note taking. The research on note taking for review has generally shown large size effects; so has research on cues to establish an intentional learning situation. The results (see Table 5) appear to be "No difference".

In trying to account for the no difference result, the investigators considered power. They had hoped for an N of over 100 since all classes had enrollments greater than 25. Subjects were told in advance that the text material was relevant to special education, that the findings would be relevant to their practice as teachers of special education and that the experimental method demonstrated was relevant to the course. (As a matter of "ethics", their criterion scores on this experiment were not to be counted for their course grade.) The absence rate was high in the first two classes. In the second two, there was a subject mutiny. Several subjects angrily stated, "We aren't going to help somebody do their dissertation," an assertion that was not factually correct. Approximately seven subjects walked out, and the morale of those remaining was dubious.

The questionnaire which accompanied the criterion test was examined for evidence of attrition among the subjects who had remained to participate in the experiment. Other research has shown that alienation can result in rebellion (e.g., walking out) or apathy (e.g., staying without interest, involvement or effort).

The questionnaire was not specifically designed to test for subject attrition. The items relevant to sub-

ject attrition are shown in Table 1. There is a reasonable basis to believe that the subjects designated by number in Table 1 were in fact non-subjects. Table 2 shows an estimated percentage of subject attrition. Almost half of the potential subjects (48%) were probably non-subjects. Given α (alpha) = .05 and σ (sigma) = 1, the actual power of the experiment was about .40 compared to the potential power of .75, which would have been the case if all the potential subjects had met the experimental control requirements.

Insert Table 1 about here

Insert Table 2 about here

The most plausible basis for the no significant difference is that the actual power is too low to detect a difference that exists. A non-experiment had taken place.

An Experiment on Lecture vs. Reading With and Without Notes (Note 3)

Fifty-eight juniors and seniors in the second course of the teacher education sequence read a 3,6000 word passage

on psycholinguistics or listened to an identical lecture. Half of each group were instructed to take notes; the other half were instructed not to. Learning was measured by a 40 item multiple choice test. The effect of reading was found to be significantly higher than listening to the same material in lecture form. The effect of taking notes was found to be significantly higher than not taking notes. The statistical analysis for this experiment is given in Tables 3 and 4.

Insert Table 3 about here.

Insert Table 4 about here.

Table 5 below shows some of the potential sources of subject attrition in this experiment. The reader can roughly determine the validity of these sources of attrition by considering whether a treatment designed for a class of normal second year college students would be effective if:

The subjects were already familiar with the treatment material.

The subjects made a minimal effort to do the treatment and/or the criterion test.

The subjects were not at all interested in learning the treatment material.

The subjects reported a very low state of arousal (e.g., sleep) during the treatment and/or the criterion test.

The subjects were in a state of apathy or depression during the treatment?

To cope with this problem, most experimenters apparently reason as follows. I recognize my experimental controls are not 100% perfect; they are not zero or constant for all subjects. However, since the subjects are randomly assigned to treatment and control groups, it is equally likely that any subjects who do not comply with or meet the requirements of the experimental controls will also be randomly distributed. Thus, my result will not be biased. If bias is suspected, additional data can be collected e.g., by questionnaire. By comparing the Xs of the treatment groups on these items, it can be established that the treatment groups do not differ on this respect, thus eliminating bias as an alternate explanation for a positive finding.

In terms of the actual power of experiment, Table 5

shows that a substantial number of different subjects, (subjects are designated by number from one to N for each of the four treatment groups) were not likely to be actual subjects.

Insert Table 5 about here.

Table 6 summarizes the data in Table 5. If subject reports are accepted as valid, 45% of potential subjects (26) have been lost as a result of attrition; thirty-two subjects remain (55%). The power of the experiment assuming α = .05 and σ = .75 substantially reduces the chances to find a difference if it exists (Actual 1 - β = .42 vs. Potential 1-β = .68).

It is also worth noting that the treatment group which heard the lecture and took notes had 10 non-subjects and only six subjects. This was the group with the highest scores, but based on this analysis, those higher scores were probably due to factors other than the treatment. Table 4 shows that subjects 1, 2, 5, 6 and 13 in the treatment group had prior learning in psycholinguistics and it may be inferred that they already knew the treatment material.

Insert Table 6 about here

As a result of re-analysis, this investigation concludes that a non-experiment has been reported in spite of the significant result.

Discussion

This article began with the psychological research paradigm in which all relevant variables which are not of direct experimental interest are held constant by direct experimental control or by random selection of subjects from the population.

If each of the subjects conforms to the requirements of each of the experimental controls, then the number of potential subjects, i.e. those who are collected or selected for an experiment, equals the number of actual subjects, i.e. those who fit the appropriate specifications and who actually comply with the experimental controls. However, for any given treatment, the wide range of individual differences makes it unlikely that the number of actual subjects equals the number of potential subjects. There is attrition. Most experimenters have as-

sumed that the attrition is random, and, therefore, findings are unbiased. However, cumulative attrition at each of a number of experimental controls can and probably does substantially reduce the actual power of an experiment. Because experimenters who calculate power in advance are likely to settle for a minimum acceptable level, even a small amount of subject attrition at each of several experimental control points can drop the actual power of an experiment below the level where a difference can be detected.

Numerous authors have commented that experiments with a large N produce significant results but are of no practical significance. Based on the rough evidence and the arguments presented here, a potential N might have to be two or more times larger than the actual N to attain an adequate actual power.

Conclusions and Recommendations

When researchers deal with large intact groups in experiments on teaching, learning and other aspects of psychology, the actual power of the experiment is likely to be substantially lower than the potential power. The assumption that randomization avoids bias of subjects who do not meet control specifications results in loss of power which in turn may result in a non-experiment. Many of the

no difference results in our literature are probably due to conducting experiments which lack sufficient actual power to detect differences that exist -- in effect non-experiments.

There is a need for explicit feedback at each point in an experiment where a variable is controlled. This can probably be accomplished by direct observation, subject questionnaires or both. The reliability of this feedback needs further research.

Researchers should calculate the actual power of an experiment from feedback on controlled variables. Where this actual power is too low, findings of no difference should be discarded. Another alternative is for the researcher to continue to collect subjects until the actual power is high enough to manifest an effect if one exists.

Non-subjects are inert subjects that contribute to within group differences; they reduce the apparent magnitude of effects tested.

If our goal is knowledge synthesis, old box score reviews must be reexamined. Where meta analysis is not possible, experiments with a low potential power should be eliminated.

The concept of non-events may be generalized to include non-experiments. Non-experiments either lack the

power to detect the difference sought from the outset or they acquire it by subject attrition. Subjects who cannot, will not or for any reason do not do the experiment are non-subjects. Synthesizing knowledge using experiments performed on large doses of non-subjects contributes to the little or nothing is known and/or the nothing makes much if any difference syndromes.

Reference Notes

1. Cronbach, L. J., & Snow, R. E. Individual differences in learning ability as a function of instructional variables. (Final Report, Contract No. OEC-4-6-061269-1217, U.S. Office of Education). Stanford, Calif.: Stanford University, School of Education, 1969.
2. Tobias, S. Research strategy on the effect of individual differences on achievement from programmed instruction. Paper presented at the meeting of the American Psychological Association, Washington, September 1969.
3. Ladas, H. Listening vs reading: With and without notes. Paper presented at the Eastern Educational Research Association Conference, Williamsburg, Virginia, March 1978.

References

Brewer, J. K. On the power of statistical tests in the American Educational Research Journal. American Educational Research Journal, 1972, 9, 391-401.

Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Charters, W. W. and Jones, J. On the risk of appraising non-events in program evaluation. Educational Researcher, 1973, 2(11), 5-7.

Cohen, J. The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 1962, 65, 145-153

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Gage, N. L. The scientific basis of the art of teaching. New York: Teachers College Press, 1978.

Guenther, W. C. Concepts of statistical inference. New York: McGraw-Hill, 1965

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., Michelson, S. Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books, 1972.

Ladas, H. Educational ethology and experiments in vacuo: Some implications for research and education, Educational Researcher, 1977, 12, 215-219

Stogdill, R. M. Handbook of leadership: A survey of theory and research. New York: The Free Press, 1974.

Trow, W. The learning process. Washington D. C.: The National Educational Association, 1954.

Table 1

Potential Sources of Subject Attrition in an Experiment on Cues vs No Cues With and Without Notes

Questionnaire Item	Treatment Groups: Cues Notes N=13	Cues No Notes N=11	No Cues Notes N=16	No Cues No Notes N=14
What percent of the time in this lecture did you spend thinking about yourself or other things not related to the lecture? 0-9 10-24 25-49 50-74 75-100 *[a]	10,[b]	8	13	1,4,5,14
To what extent did you make an effort to learn from the lecture? minimal* sub-par average moderate maximum	10	11	1,7,8, 10,14	1,4,14
Were you interested in learning this material on psycho-linguistics? not-at-all* slightly somewhat very much	5,	1,8,11	15	1,2,4
Was the volume of the lecturer's voice... poor* adequate good	8		6	
was the background noise low medium high*				3
What was your interest level in this topic? very low* low average high very high	2,5,6,10	1,3,8,11	8,14,15	3,4,5,8
How complete are your notes? very inadequate* somewhat less than adequate adequate somewhat more than adequate very adequate	13	6		

Notes: a. Starred answer indicates option for which attrition is assumed.

b. Subjects are numbered 1 → N in each treatment group.

Table 2

Number and Percent of Non Subjects and Actual Subjects

Based on Table 1

Treatment Group	Potential Subjects		Non Subjects		Actual Subjects	
	N	%	N	%	N	%
Cues						
Notes	13		6		7	
No Notes	11		5		6	
No Cues						
Notes	16		8		8	
No Notes	14		7		7	
Total	54	100	26	48	28	52

Table 3

Immediate Retention Test Scores and Time To Complete Treatment

Means, Standard Deviation, and Number of Subjects

Group	Score			Time to complete[a]		
	M	SD	N	M	SD	N
RN	26.50	3.11	14	27.83	3.95	12
RnN	21.29	4.57	14	22.69	4.13	13
LN	20.44	5.18	16	26.50	1.55[b]	16
LnN	17.86	5.87	14	26.71	1.54	14
Totals	21.48	5.62	58	25.95	3.43	55[c]

NOTE. Abbreviations: R = Read; L = Lecture; N = Notes; nN = no Notes.

[a]In minutes.

[b]In one section the lecture took 25 minutes (144wpm) in the other 28 minutes (129wpm).

[c]No data on 3 subjects.

Table 4

Analysis of Variance of Immediate Retention Scores

Source of Variation	df	Mean Square	F
Read vs. Lecture	1	331.65	14.37 *
Notes vs. No Notes	1	214.97	9.32 *
Interaction	1	25.07	1.09
Within groups	54	23.04	

* $p < .01$

Table 5

Potential Sources of Subject Attrition in an Experiment on Lecture vs. Reading With and Without Notes

Questionnaire Item	Treatment Groups: Lecture Notes N=16	Lecture No Notes N=14	Reading Notes N=14	Reading No Notes N=14
Have you taken a course in linguistics?[a] (Already familiar?) yes* no	1,2,5,6, 13	8	4,8	2,4,7,8, 11,13
To what extent did you make an effort to learn from the lecture or reading? minimal* sub-par average moderate maximum	2,6,11, 14,16	5,8		1,
To what extent did you make an effort to do the test? minimal* sub-par average moderate maximum	6,11,14			
Do you take notes during lectures in other courses? never* occasionally often usually always	8			
Were you interested in learning this material on psycholinguistics? not at all* slightly somewhat very much	10,14	5	11	
What was your average level of arousal during the lecture or reading? very low* slightly low optimum slightly high very high	6	9		
What was your average level of arousal during the test? very low* slightly low optimum slightly high very high	6			
What was your emotion during the lecture or reading?	8,14	5	2	1
apathy* depression grief sadness embarassment shame,regret fear anxiety anger irritation antagonism boredom cheerfulness serenity enthusiasm				
What was your emotion during the test? (Same possible answers as item above.)			9	1
Did you take notes? yes no*	1		9	
If you took notes, how complete were they? very inadequate * somewhat less than adequate adequate somewhat more than adequate very adequate	5,6,8, 11, 14		5	

(Table 5 continued)

Notes: a. The material was on psycholinguistics so any subjects who had taken it may have already been familiar with the treatment.

b. The starred answers (*) represents potential attrition.

c. Subjects are designated by numbers from 1 to N for each treatment group.

d. Arousal level was presented with a curve and examples on the curve eg. very low arousal was labeled "eg. sleep"

Table 6

Number and Percent of Non Subjects and Actual Subjects

Based on Table 1

Treatment Group	Potential Subjects		Non Subjects		Actual Subjects	
	N	%	N	%	N	%
Lecture						
Notes	16		10		6	
No Notes	14		3		11	
Reading						
Notes	14		6		8	
No Notes	14		7		7	
Total	58	100	26	45	32	55